Bernoulli **11**(6), 2005, 1031–1057

Profile likelihood inferences on semiparametric varying-coefficient partially linear models

JIANQING FAN[1] and TAO HUANG[2]

[1]*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA. E-mail: jqfan@princeton.edu*
[2]*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06511, USA. E-mail: t.huang@yale.edu*

Varying-coefficient partially linear models are frequently used in statistical modelling, but their estimation and inference have not been systematically studied. This paper proposes a profile least-squares technique for estimating the parametric component and studies the asymptotic normality of the profile least-squares estimator. The main focus is the examination of whether the generalized likelihood technique developed by Fan *et al.* is applicable to the testing problem for the parametric component of semiparametric models. We introduce the profile likelihood ratio test and demonstrate that it follows an asymptotically χ^2 distribution under the null hypothesis. This not only unveils a new Wilks type of phenomenon, but also provides a simple and useful method for semiparametric inferences. In addition, the Wald statistic for semiparametric models is introduced and demonstrated to possess a sampling property similar to the profile likelihood ratio statistic. A new and simple bandwidth selection technique is proposed for semiparametric inferences on partially linear models, and numerical examples are presented to illustrate the proposed methods.

Keywords: generalized likelihood ratio statistics; local linear regression; partially linear models; profile likelihood; varying-coefficient partially linear models; wald statistics

1. Introduction

With the improvement of computing facilities over the last three decades, there has been an upsurge of interest and effort in nonparametric models as researchers have realized that parametric models are inadequate in capturing the relationship between the response variable and its associated covariates in many practical situations. Such data-analytic approaches are useful for exploring the hidden structure but can be too flexible to draw concise conclusions. For an introduction to nonparametric techniques, see Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995) and Fan and Gijbels (1996), among others. Even though many useful techniques have been proposed, the tools available for inferences on semiparametric and nonparametric models are limited.

In an effort to derive a generally applicable testing approach, Fan *et al.* (2001) proposed the generalized likelihood ratio (GLR) statistic for nonparametric models. Their motivation

1350–7265

was as follows. The maximum likelihood ratio test statistic in general may not exist in nonparametric and semiparametric settings. Even if it does, it is hard to find and may not be optimal in the simplest nonparametric regression setting. These drawbacks can be avoided when the maximum likelihood estimator is replaced by other reasonable nonparametric estimators, resulting in a class of statistics called the GLR statistic. The GLR test is intuitively appealing. Fan *et al.* (2001) showed that for a variety of models and a number of nonparametric versus nonparametric and parametric versus nonparametric testing problems, the null distribution of the GLR test statistic follows an asymptotically χ^2 distribution, independent of nuisance parameters. This property is called the Wilks phenomenon and facilitates the application of the GLR statistic. The critical value can be determined either by asymptotic distributions or by simulations. Furthermore, Fan *et al.* (2001) showed that the GLR test is asymptotically optimal in the sense of Ingster (1993). The question arises naturally whether the generalized likelihood technique is applicable to semiparametric models. This forms the main theme of the present study.

Like nonparametric models, semiparametric models have various forms. A useful semiparametric model that facilitates the study of the GLR test, or, more specifically, the profile likelihood ratio (PLR) test, is the varying-coefficient partially linear model. This is an extension of the varying-coefficient model, and has recently been studied by Zhang *et al.* (2002) and Li *et al.* (2002).

Let Y be the response variable and $(U, \mathbf{X}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})$ be its associated covariates. The varying-coefficient partially linear model assumes the following structure:

$$Y = \boldsymbol{\alpha}^{\mathrm{T}}(U)\mathbf{X} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \varepsilon, \tag{1.1}$$

where ε is independent of $(U, \mathbf{X}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})$ and has $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = \sigma^2$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ is a q-dimensional vector of unknown parameters and $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_p(\cdot))^{\mathrm{T}}$ is a p-dimensional vector of unknown coefficient functions. Due to the curse of dimensionality, we assume, for simplicity, that U is univariate. Model (1.1) allows interaction between the covariates U and $\mathbf{X}$ in such a way that a different level of covariate U is associated with a different linear model. This enables us to examine the extent to which the effects of covariates $\mathbf{X}$ vary over different levels of the covariate U.

The varying-coefficient model arises in many different contexts and has been successfully applied to multidimensional nonparametric regression, generalized linear models, time series analysis, longitudinal and functional data analysis, and time-varying models in finance. Early applications of the varying-coefficient model appeared in Haggan and Ozaki (1981) in the time series context. However, nonparametric techniques were not popularized until the work of Cleveland *et al*. (1991), Chen and Tsay (1993), and Hastie and Tibshirani (1993). For nonparametric regression models, Carroll *et al*. (1998) proposed a method that is based on the local estimation equations, and Fan and Zhang (1999) proposed a two-step procedure to accommodate varying degrees of smoothness among coefficient functions. Xia and Li (1999) derived the distribution of the maximum discrepancy between the estimated coefficients and their true values, and Cai *et al*. (2000) applied the varying-coefficient techniques to the generalized linear model. The varying-coefficient model has also been popularly used to analyse longitudinal data. It allows one to examine the extent to which

the association between response and covariates varies over time. See, for example, the work of Brumback and Rice (1998), Hoover *et al*. (1998), and Huang *et al*. (2002).

When $p = 1$ and $\mathbf{X} = 1$, (1.1) becomes partially linear model. It has been widely studied in the literature. See, for example, the work of Wahba (1984), Cuzick (1992), and Severini and Wong (1992). Speckman (1988) introduced the idea of profile least-squares for the partially linear model. Liang *et al.* (1999) studied the partial linear model with errors-in-variables. Härdle *et al*. (1998) investigated the problem of testing linearity in the nonparametric component. More references and techniques can be found in the recent monograph by Härdle *et al*. (2000).

The parametric component in semiparametric models is frequently of primary interest. It has explanatory power that is similar to parametric models. It is natural to investigate whether certain variables in the parametric component are statistically significant after fitting the model. This leads to testing problems such as

$$H_0 : \beta_1 = \ldots = \beta_l = 0, \qquad l \leqslant q.$$

More generally, one may consider the linear hypothesis

$$H_0 : \ \mathbf{A}\boldsymbol{\beta} = \mathbf{0}, \tag{1.2}$$

where $\mathbf{A}$ is a given $l \times q$ full rank matrix. Here one is testing a semiparametric hypothesis versus another semiparametric hypothesis. The conventional maximum likelihood ratio test cannot be applied, because the nonparametric maximum likelihood estimates do not exist for functions $\boldsymbol{\alpha}(\cdot)$. A natural alternative is to relax the requirement on the estimates of functions $\boldsymbol{\alpha}(\cdot)$ to be any reasonable nonparametric estimates and use them to construct the likelihood ratio test. This yields a family of test statistics known as the GLR statistic. As the nonparametric estimator will be constructed by using the profile likelihood technique, the resulting test statistic is more specifically referred to as the profile likelihood ratio (PLR) statistic. The question is, then, whether the Wilks type result holds and whether the traditional power calculation continues to apply.

We first introduce profile least-squares (normal likelihood) estimation and then establish the asymptotic normality for the profile least-squares estimate. With the asymptotic normality result, one can easily construct an estimated covariance matrix and hence the Wald statistic. In contrast, one can proceed directly to construct the PLR statistic without referring to the estimated covariance matrix. This is an advantage of the PLR test. It turns out that the PLR test for the parametric component behaves very much like the maximum likelihood ratio test for parametric models, even though the profile likelihood estimator is not the maximum likelihood estimator.

This paper is organized as follows. Section 2 introduces profile least-squares estimation. In Sections 3 and 4 we construct the PLR and Wald statistics. The asymptotic distributions of both statistics are derived under regularity conditions. In Section 5 we briefly address the issue of inferences on the nonparametric component of semiparametric models, and in Section 6 we propose a simple and effective bandwidth selection method for the partially linear model. Section 7 contains several numerical results. Technical proofs are relegated to the Appendix.

2. Profile least-squares estimation

There are many approaches to estimating the unknown parameters $\{\beta_j, j = 1, \ldots, q\}$ and the varying coefficient functions $\{\alpha_i(\cdot), i = 1, \ldots, p\}$. Profile least squares is a useful approach and will be shown to be semiparametrically efficient for model (1.1). When $\varepsilon \sim N(0, \sigma^2)$, the approach becomes profile likelihood estimation; see, for example, Speckman (1988), Severini and Wong (1992) and Carroll *et al.* (1997).

Suppose that we have a random sample of size n, $\{(U_k, X_{k1}, \ldots, X_{kp}, Z_{k1}, \ldots, Z_{kq}, Y_k), k = 1, \ldots, n\}$, from model (1.1). For any given $\boldsymbol{\beta}$, (1.1) can be written as

$$Y_k^* = \sum_{i=1}^{p} \alpha_i(U_k) X_{ki} + \varepsilon_k, \qquad k = 1, \ldots, n, \tag{2.1}$$

where $Y_k^* = Y_k - \sum_{j=1}^{q} \beta_j Z_{kj}$. This transforms the varying-coefficient partially linear model (1.1) into the varying-coefficient model (2.1). The local linear regression technique is applied to estimate the coefficient functions $\{\alpha_i(\cdot), i = 1, \ldots, p\}$ in (2.1). For u in a small neighbourhood of u_0, one can approximate $\alpha_i(u)$ locally by a linear function

$$\alpha_i(u) \approx \alpha_i(u_0) + \alpha_i'(u_0)(u - u_0) \equiv a_i + b_i(u - u_0), \qquad i = 1, \ldots, p.$$

This leads to the following weighted local least-squares problem: find $\{(a_i, b_i), i = 1, \ldots, p\}$ so as to minimize

$$\sum_{k=1}^{n} \left[Y_k^* - \sum_{i=1}^{p} \{a_i + b_i(U_k - u_0)\} X_{ki} \right]^2 K_h(U_k - u_0), \tag{2.2}$$

where K is a kernel function, h is a bandwidth and $K_h(\cdot) = K(\cdot/h)/h$.

Let us work with the matrix notation. Denote $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$, $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iq})^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$, $\mathbf{W}_u = \mathrm{diag}(K_h(U_1 - u), \ldots, K_h(U_n - u))$, and

$$\mathbf{M} = \begin{pmatrix} \boldsymbol{\alpha}^{\mathrm{T}}(U_1)\mathbf{X}_1 \\ \vdots \\ \boldsymbol{\alpha}^{\mathrm{T}}(U_n)\mathbf{X}_n \end{pmatrix}, \qquad \mathbf{D}_u = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}} & \dfrac{U_1 - u}{h}\mathbf{X}_1^{\mathrm{T}} \\ \vdots & \vdots \\ \mathbf{X}_n^{\mathrm{T}} & \dfrac{U_n - u}{h}\mathbf{X}_n^{\mathrm{T}} \end{pmatrix}.$$

Then (2.1) can be written as

$$\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta} = \mathbf{M} + \boldsymbol{\varepsilon}. \tag{2.3}$$

The solution to the problem (2.2) is given by

$$[\hat{a}_1(u), \ldots, \hat{a}_p(u), h\hat{b}_1(u), \ldots, h\hat{b}_p(u)]^{\mathrm{T}} = \{\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{D}_u\}^{-1}\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}).$$

The estimator for $\mathbf{M}$ is then

$$\hat{\mathbf{M}} = \begin{pmatrix} [\mathbf{X}_1^{\mathrm{T}} & 0]\{\mathbf{D}_{u_1}^{\mathrm{T}}\mathbf{W}_{u_1}\mathbf{D}_{u_1}\}^{-1}\mathbf{D}_{u_1}^{\mathrm{T}}\mathbf{W}_{u_1} \\ \vdots \\ [\mathbf{X}_n^{\mathrm{T}} & 0]\{\mathbf{D}_{u_n}^{\mathrm{T}}\mathbf{W}_{u_n}\mathbf{D}_{u_n}\}^{-1}\mathbf{D}_{u_n}^{\mathrm{T}}\mathbf{W}_{u_n} \end{pmatrix} (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}) = \mathbf{S}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}). \tag{2.4}$$

The matrix $\mathbf{S}$ is a smoothing matrix and depends only on the observations $\{(U_i, \mathbf{X}_i^{\mathrm{T}}), i = 1, \ldots, n\}$. Substituting $\hat{\mathbf{M}}$ into (2.3), we obtain

$$(\mathbf{I} - \mathbf{S})\mathbf{Y} = (\mathbf{I} - \mathbf{S})\mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.5}$$

Applying least squares to the linear model (2.5), we obtain

$$\hat{\boldsymbol{\beta}} = \{\mathbf{Z}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\mathbf{Z}\}^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\mathbf{Y}. \tag{2.6}$$

Moreover,

$$\hat{\mathbf{M}} = \mathbf{S}(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}). \tag{2.7}$$

In practice, some of the U_i can fall in the sparse regions where their local neighbourhoods contain only a few data points. The functions $\boldsymbol{\alpha}(\cdot)$ cannot be estimated well at these U_i, and hence their corresponding rows should be eliminated from $\mathbf{S}$ in order not to adversely affect the estimation of $\boldsymbol{\beta}$. To facilitate the notation, we keep all the rows and use the assumption that U has a bounded support with a non-vanishing density on its support to avoid the sparsity problem.

3. Profile likelihood ratio test

3.1. PLR statistic

To gain the insight into the construction of the PLR statistic, assume for the moment that $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$. As demonstrated in Fan *et al.* (2001) and by our proof in the Appendix, the normality assumption is used merely to motivate the procedure. Under model (1.1), the likelihood function is given by

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) = -n\log(\sqrt{2\pi}\sigma) - \frac{\mathrm{RSS}_1}{2\sigma^2},$$

where $\mathrm{RSS}_1 = \sum_{i=1}^n (Y_i - \boldsymbol{\alpha}(U_i)^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i)^2$. For a given $\boldsymbol{\beta}$, $\boldsymbol{\alpha}(\cdot)$ is estimated by the local linear fit and results in the estimator $\hat{\boldsymbol{\alpha}}(\cdot; \boldsymbol{\beta})$. Substituting this into the above likelihood function, we obtain

$$\ell(\hat{\boldsymbol{\alpha}}(\cdot; \boldsymbol{\beta}), \boldsymbol{\beta}, \sigma) = -n\log(\sqrt{2\pi}\sigma) - \frac{\mathrm{RSS}_1}{2\sigma^2}, \tag{3.1}$$

where, with a slight abuse of notation, $\mathrm{RSS}_1 = \sum_{i=1}^n (Y_i - \hat{\boldsymbol{\alpha}}(U_i, \boldsymbol{\beta})^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i)^2$. Maximizing (3.1) with respect to $\boldsymbol{\beta}$ and σ^2 yields the profile likelihood estimator $\hat{\boldsymbol{\beta}}$ given by (2.6) and

$$\hat{\sigma}^2 = n^{-1}\mathrm{RSS}_1,$$

where $\mathrm{RSS}_1 = \sum_{i=1}^n (Y_i - \hat{\mathbf{M}}_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{Z}_i)^2$. Note that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are not maximum likelihood

estimators because $\hat{\boldsymbol{\alpha}}(\cdot;\boldsymbol{\beta})$ is not obtained by the maximum likelihood method. Substituting these estimates into (3.1) yields the profile likelihood

$$\ell(H_1) = -\frac{n}{2}\log\left(\frac{2\pi}{n}\right) - \frac{n}{2}\log(\mathrm{RSS}_1) - \frac{n}{2}.$$

To facilitate the notation, write $\widetilde{\mathbf{Y}} = (\mathbf{I}-\mathbf{S})\mathbf{Y}$ and $\widetilde{\mathbf{Z}} = (\mathbf{I}-\mathbf{S})\mathbf{Z}$. Now, the profile likelihood estimator can simply be written as

$$\hat{\boldsymbol{\beta}} = \{\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}}\}^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Y}} \quad \text{and} \quad \hat{\mathbf{M}} = \mathbf{S}(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\beta}}).$$

On the other hand, under the null hypothesis (1.2), the profile likelihood estimator is the one that maximizes (3.1) subject to constraint (1.2). The solution is the same as for the constrained least-squares problem for the 'synthetic linear model' (2.5), and is given by

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} - (\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\mathbf{A}^{\mathrm{T}}\{\mathbf{A}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\mathbf{A}^{\mathrm{T}}\}^{-1}\mathbf{A}\hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\mathbf{M}}_0 = \mathbf{S}(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\beta}}_0).$$

Hence, under the null hypothesis, maximizing (3.1) with respect to $\boldsymbol{\beta}$ and σ^2 yields the profile likelihood

$$\ell(H_0) = -\frac{n}{2}\log\left(\frac{2\pi}{n}\right) - \left(\frac{n}{2}\right)\log(\mathrm{RSS}_0) - \frac{n}{2},$$

where $\mathrm{RSS}_0 = \sum_{i=1}^{n}(Y_i - \hat{\mathbf{M}}_{0i} - \hat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{Z}_i)^2$.

Now, with the profile likelihood derived above, the GLR statistic is constructed as

$$T_n = \ell_n(H_1) - \ell_n(H_0) = \frac{n}{2}\log\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1} \approx \frac{n}{2}\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}. \tag{3.2}$$

This specific form is referred to as the PLR statistic. First, as noted above, the PLR test is not the same as the maximum likelihood ratio test. Second, the same bandwidth is used for constructing profile likelihood estimators under both null and alternative hypotheses, which is the key to the success of the method. Third, as the maximum likelihood ratio test in the parametric model, the PLR test can simply be formed without referring to the standard error formula for $\hat{\boldsymbol{\beta}}$, which is one advantage of the method.

3.2. Wilks phenomenon

The PLR statistic is derived analogously to the maximum likelihood ratio statistic. However, they are also very different. The nuisance functions $\{\alpha_i(\cdot), i = 1, \ldots, p\}$ are fully nonparametric and hence the parameter space is infinite-dimensional. They are not estimated by the maximum likelihood method. The question then arises as to whether the asymptotic null distribution of the PLR statistic is still χ^2. The following theorem shows that the traditional likelihood theory continues to apply.

Theorem 3.1. *Under the null hypothesis* (1.2) *and the conditions in the Appendix, the PLR statistic* $2T_n(h)$ *follows the asymptotic* χ^2 *distribution with* l *degrees of freedom.*

Theorem 3.1 shows that the asymptotic null distribution of $2T_n(h)$ is independent of the design and the nuisance parameters σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}(\cdot)$, and is χ^2 with l degrees of freedom for testing (1.2). Because of this, the critical value can be computed either by asymptotic distributions or by simulations with nuisance parameter values taken to be reasonable estimates. These estimates should be obtained from the full model, namely model (1.1) without assumption (1.2).

Theorem 3.1 not only provides a useful result for testing the parametric component in semiparametric models, but also unveils a new type of Wilks phenomenon for the GLR test. Fan *et al.* (2001) showed that the Wilks type result holds for testing the parametric versus nonparametric or nonparametric versus nonparametric type of hypotheses. However, whether the nonparametric estimates are accurate enough for constructing the likelihood ratio test for the parametric component remains unknown. The result in Theorem 3.2 gives a definite answer; it allows one to proceed to the likelihood ratio test as if the model were parametric. It also sheds light on directions for future research on semiparametric inferences.

The key to the Wilks phenomenon is the orthogonality of score functions, which is inherited from the profile least-squares method. This condition also makes the profile least-squares method semiparametrically efficient. See Severini and Wong (1992) and Theorem 4.1 below for details. When other estimators are used, the orthogonality condition may fail and the GLR tests may not possess the Wilks phenomenon. An example of this is given in Härdle *el al.* (2004) where the marginal integration is used for constructing nonparametric components in a partially linear additive model. In addition, as demonstrated in Härdle *et al*. (1998), the Wilks phenomenon does not hold when the profile least-squares estimates are used in a different way.

3.3. Power of PLR test

We now provide the formula for the calculation of the power of the PLR test under the contiguous alternatives, where $\mathbf{A}\boldsymbol{\beta}$ converges to zero at the root n rate. The power formula enables us to not only determine the sample size for semiparametric testing problems, but also compare the power with the Wald test, as presented in the next section.

Theorem 3.2. *Under the alternative hypothesis to problem* (1.2) *and the conditions in the Appendix, the PLR statistic* $2T_n(h)$ *follows the asymptotic non-central* χ^2 *distribution with* l *degrees of freedom, and non-centrality parameter* $\lambda = \lim_{n\to\infty} n\boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\{\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}\}^{-1}\mathbf{A}\boldsymbol{\beta}$, *where*

$$\boldsymbol{\Sigma} = \sigma^2\{\mathrm{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) - \mathrm{E}\big[\mathrm{E}(\mathbf{Z}\mathbf{X}^{\mathrm{T}}|U)\mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U)^{-1}\mathrm{E}(\mathbf{X}\mathbf{Z}^{\mathrm{T}}|U)\big]\}^{-1}.$$

In the computation of power for a given value of $\boldsymbol{\beta}$, we need an estimate of $\boldsymbol{\Sigma}$. This will be given by (4.2) or (4.3) below.

4. Wald test

The testing problem (1.2) can also be handled by using the Wald statistic,

$$W_n(h) = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\hat{\boldsymbol{\Sigma}}_h\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\hat{\boldsymbol{\beta}}, \tag{4.1}$$

where $\hat{\boldsymbol{\Sigma}}_h$ is an estimated covariance matrix for $\boldsymbol{\beta}$, which involves estimated functions $\hat{\boldsymbol{\alpha}}(\cdot)$ and depends on a smoothing parameter h. Following (2.6), an estimate of the covariance matrix for $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\Sigma}}_h = n\hat{\sigma}^2(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I}-\mathbf{S})(\mathbf{I}-\mathbf{S})^{\mathrm{T}}\widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}, \tag{4.2}$$

where $\hat{\sigma}^2$ is the sample variance of residuals. A simpler estimate of the covariance matrix is

$$\hat{\boldsymbol{\Sigma}}_h^* = n\hat{\sigma}^2(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}. \tag{4.3}$$

To facilitate the presentation, we will not discuss the estimation of σ^2. In Lemmas A.2 and A.3, we will show that both $\hat{\boldsymbol{\Sigma}}_h$ and $\hat{\boldsymbol{\Sigma}}_h^*$ are consistent estimates of $\boldsymbol{\Sigma}$. The PLR statistic has the advantage that it can be formed without reference to the estimated covariance matrix for $\boldsymbol{\beta}$.

4.1. Asymptotic normality

Theorem 4.1. *Under the conditions in the Appendix, the profile likelihood estimator of* $\boldsymbol{\beta}$ *is asymptotically normal, that is,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ *is given in Theorem 3.2.*

Consider the partially linear model, where $p = 1$ and $\mathbf{X} \equiv 1$. Then

$$\mathrm{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) - \mathrm{E}\big[\mathrm{E}(\mathbf{Z}|U)\mathrm{E}(\mathbf{Z}^{\mathrm{T}}|U)\big] = \mathrm{E}\{\mathrm{var}(\mathbf{Z}|U)\},$$

and Theorem 4.1 is consistent with the result of Carroll *et al.* (1997). In fact, they showed that $\boldsymbol{\Sigma}$ is the semiparametric information bound (see Bickel *et al.* 1993). One can follow the formulation and the results of Chamberlain (1992) to show that $\boldsymbol{\Sigma}$ is a semiparametric efficient bound for the general varying-coefficient partially linear model. Hence, the profile likelihood estimator is semiparametrically efficient.

4.2. The null distribution and power

The Wald statistic (4.1), based on the semiparametrically efficient estimator $\hat{\boldsymbol{\beta}}$, is intuitively appealing and serves as a benchmark for other procedures. The following two theorems give the asymptotic null distribution and the power.

Theorem 4.2. *Under the null hypothesis* (1.2) *and the conditions in the Appendix, the Wald statistic* $W_n(h)$ *follows the asymptotic* χ^2 *distribution with* l *degrees of freedom.*

Theorem 4.3. *Under the alternative hypothesis to the problem* (1.2) *and the conditions in*

the Appendix, the Wald statistic $W_n(h)$ follows the asymptotic non-central χ^2 distribution with l degrees of freedom, and non-centrality parameter $\lambda = \lim_{n\to\infty} n\boldsymbol{\beta}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\boldsymbol{\beta}$.

Theorems 4.2 and 4.3 show that the Wald statistic possesses the same asymptotic null and alternative distributions as the PLR statistic, which gives theoretical endorsement to the PLR statistic. On the other hand, the PLR statistic is simpler to use for many applications.

5. Inferences on nonparametric component

After obtaining nonparametric estimates of $\{\boldsymbol{\alpha}_1(\cdot), \ldots, \boldsymbol{\alpha}_p(\cdot)\}$, researchers frequently ask whether certain parametric models fit the nonparametric components. This leads us to consider hypothesis testing problems such as:

$$H_0 : \alpha_p(U) = \alpha_p(U, \boldsymbol{\theta}) \quad \text{versus} \quad H_1 : \alpha_p(U) \neq \alpha_p(U, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ is an unknown vector. The above problem includes testing the significance of the variable X_p in which $a_p(\cdot; \boldsymbol{\theta}) = 0$ and the homogeneity of the model in which $a_p(\cdot; \boldsymbol{\theta}) = a_p$. Other testing problems, such as the cases where only some components have constant or zero coefficients, can be similarly dealt with. The essence is that the parametric component can be estimated at $O(n^{-1/2})$-consistency and regarded as known in nonparametric inferences. Hence, the techniques and the results of Fan *et al.* (2001) can be extended to the parametric component in the semiparametric model. As an illustration, we consider the problem of testing the homogeneity:

$$H_0 : \alpha_1(\cdot) = \alpha_1, \ldots, \alpha_p(\cdot) = \alpha_p. \tag{5.1}$$

Let $\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_p$ and $\widetilde{\boldsymbol{\beta}}$ be the least-squares estimators under H_0. Following the derivation of Fan *et al.* (2001), the GLR statistic is defined as

$$T_0 = \frac{n}{2}\log\frac{\mathrm{RSS}(H_0)}{\mathrm{RSS}(H_1)},$$

where $\mathrm{RSS}(H_0) = \sum_{i=1}^n (Y_i - \sum_{j=1}^p \widetilde{\alpha}_j X_{ij} - \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{Z}_i)^2$, and $\mathrm{RSS}(H_1) = \sum_{i=1}^n (Y_i - \sum_{j=1}^p \hat{\alpha}_j(U_i)X_{ij} - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{Z}_i)^2$. The following result is an extension of Theorem 5 of Fan *et al.* (2001).

Theorem 5.1. *Under the null hypothesis* (5.1) *and the conditions in the Appendix, if $h \to 0$ in such a way that $nh^{3/2} \to \infty$, then the GLR statistic*

$$r_K T_0 \overset{a}{\sim} \chi^2_{\delta_n},$$

where $|\boldsymbol{\Omega}|$ is the length of the support of U and

$$r_K = \frac{K(0) - \frac{1}{2}\int K^2(t)\mathrm{d}t}{\int\left(K(t) - \frac{1}{2}K * K(t)\right)^2 \mathrm{d}t} \quad \text{and} \quad \delta_n = r_K \frac{p|\boldsymbol{\Omega}|}{h}\left(K(0) - \frac{1}{2}\int K^2(t)\mathrm{d}t\right).$$

As discussed above, the asymptotic distribution of the nonparametric component can be obtained from that in the varying-coefficient model because the parametric component can be regarded as known. See, for example, Fan and Zhang (1999; 2001), Xia and Li (1999) and Zhang *et al*. (2002).

6. Implementation for partially linear model

The profile likelihood estimator depends on the choice of bandwidth. Furthermore, the PLR and Wald statistics also involve the choice of bandwidth. The issue of bandwidth selection arises naturally in practice. The issue of selecting bandwidths for semiparametric models, particularly for estimating the parametric component, was posed by Bickel and Kwon (2001) as an important and unsolved problem. First, we would like to observe that the performance of the profile least-squares estimate $\hat{\boldsymbol{\beta}}$ and the PLR and Wald statistics does depend not very sensitively on the choice of bandwidth, as long as h is not too large to create excessive bias; see condition (6) in the Appendix and the discussion by Fan in Bickel and Kwon (2001). The reason is that the bias in the estimation of the nonparametric component cannot be averaged out in the process of estimating the parametric component, but the variance can be averaged out.

6.1. Bandwidth selection for partially linear model

When $p = 1$ and $\mathbf{X} \equiv 1$, model (1.1) becomes the partially linear model,

$$Y = \boldsymbol{\alpha}(U) + \sum_{j=1}^{q} \beta_j Z_j + \varepsilon. \tag{6.1}$$

Let $\{(U_i, Z_{i1}, \ldots, Z_{iq}, Y_i), i = 1, \ldots, n\}$ be a random sample of size n from the partially linear model (6.1), ordered according to the variable U. Under some mild conditions, the spacing between U_{i+1} and U_i is $O_P(1/n)$ so that $\alpha(U_{i+1}) - \alpha(U_i) = O_p(1/n)$. Then, by model (6.1),

$$\begin{aligned} Y_{i+1} - Y_i &= \alpha(U_{i+1}) - \alpha(U_i) + \beta_1(Z_{i+1,1} - Z_{i,1}) + \ldots + \beta_q(Z_{i+1,q} - Z_{i,q}) + \varepsilon_{i+1} - \varepsilon_i \\ &\approx \gamma_0 + \gamma_1(U_{i+1} - U_i) + \beta_1(Z_{i+1,1} - Z_{i,1}) + \cdots + \beta_q(Z_{i+1,q} - Z_{i,q}) + \varepsilon_i^*, \end{aligned} \tag{6.2}$$

where ε_i^* are correlated stochastic errors with $\varepsilon_i^* = \varepsilon_{i+1} - \varepsilon_i$. Thus, the nonparametric function $\alpha(\cdot)$ in the partially linear model (6.1) is eliminated. The coefficients $\gamma_0, \gamma_1, \beta_1, \ldots, \beta_q$ can be estimated using ordinary least squares from the approximated linear model (6.2). This kind of idea appears independently in the work of Yatchew (1997), who used $\gamma_0 = \gamma_1 = 0$, and Fan and Huang (2001), who used the linear terms to gain a better approximation. Set

$$Y_i^* = Y_i - \hat{\beta}_1^0 Z_{i,1} - \cdots - \hat{\beta}_q^0 Z_{i,q},$$

where $\hat{\beta}_1^0, \ldots, \hat{\beta}_q^0$ are estimated from (6.2). We call such an estimate the difference-based

estimate (DBE). Then, $Y_i^* \approx \alpha(U_i) + \varepsilon_i^*$, which is a univariate nonparametric regression problem. Therefore, one can apply univariate bandwidth selection procedures such as the preasymptotic substitution method (Fan and Gijbels 1995), the plug-in bandwidth selector (Ruppert *et al*. 1995), and the empirical bias method (Ruppert 1997) to select a smoothing parameter h. In this paper, we use the empirical bias method to choose the bandwidth for semiparametric estimation and inference.

6.2. A simple F-test for partially linear model

For the approximated linear model (6.2), problem (1.2) becomes a linear hypothesis, and the F-statistic can be employed. Note that the noise terms $\{\varepsilon_i^*, i = 1, \ldots, n-1\}$ are no longer independent of each other. Hence, the F-statistic will not have a correct null distribution.

To avoid the aforementioned dependence problem, we consider a simpler version of the approximated linear model (6.2). Instead of using all $\{Y_{i+1} - Y_i\}$, we use $\{Y_{2i+1} - Y_{2i}\}$ to construct the approximated linear model (6.2). By doing so, the independence of the data is inherited. We lose the data $\{Y_{2i+1} + Y_{2i}\}$, which contain less information about $\boldsymbol{\beta}$ than $\{Y_{2i+1} - Y_{2i}\}$, as the former contains the nuisance function $\boldsymbol{\alpha}(\cdot)$ while the latter does not. Thus, the efficiency based on the model (6.2) with the selected subsample should, intuitively, be at least 50%. In the next section, we compare the efficiency of such an F-test with the more sophisticated PLR and Wald tests via simulations.

7. Numerical studies

In this section, we present the results of three Monte Carlo simulations to show the finite-sample performance of the profile least-squares estimate and three proposed testing procedures. Throughout this section, we use the Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$.

7.1. A partially linear model

Consider the partially linear model

$$Y = \sin(2U) + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \varepsilon,$$

where U, Z_1, Z_2, Z_3 are covariates. The covariates U, Z_1, and Z_2 are jointly normally distributed with mean 0 and variance 1. Furthermore, the correlation coefficients among these three random variables are all 0.5. The covariate Z_3 is binary, independent of U, Z_1 and Z_2, taking the value 1 with probability 0.4 and the value 0 with probability 0.6. The true parameter for β_1 is always fixed at $\beta_1 = 2$, and β_2 and β_3 are taken differently for different problems. To gain an idea of the effect of the normality assumption on our results, we consider two cases: (i) $\varepsilon \sim N(0, 1)$ and (ii) $\varepsilon \sim \frac{2}{3}N(0, \frac{1}{2}) + \frac{1}{3}N(0, 2)$, a mixture normal with mean 0 and variance 1. For each simulation, we draw 1000 random samples of size 100 from the above model and employ the bandwidth selection scheme in Section 6.1. The true parameters are taken as $\beta_1 = 2$, $\beta_2 = 2\theta$ and $\beta_3 = \theta$ with $\theta = 0$ and $\theta = 0.5$. For

constructing the profile least-squares estimate for the parametric component, we only use those estimates with $|U_i| \leqslant 1.645$, which basically rules out approximately 10% of data.

The first aim of this simulation study is to show that the performance of the profile least-squares estimator does not depend sensitively on the choice of bandwidth. To demonstrate this, we fix the smoothing parameter at three values $h = 0.625/1.5$, 0.625 and 0.625×1.5. Note that the optimal bandwidth for estimating the nonparametric component is about $h = 0.625$. We also report the performance of the DBE. The mean and standard deviation based on 1000 simulations are reported in Table 1 for case (i).

The second aim of this simulation study is to examine the accuracy of the standard error formula given by (4.2). The 1000 estimated standard errors from 1000 simulations are summarized by its mean estimated value (denoted by *SD*) and its standard deviation (denoted by SD_{std}), which are also reported in Table 1. The 1000 estimated standard errors are surprisingly close to the standard deviation of 1000 estimated coefficients (denoted by SD_{m}). The latter can be regarded as the true standard error of an estimation procedure, which shows that our standard error formula is very accurate.

The third aim of this simulation is to study the performance of the proposed testing methods. We consider the null hypothesis $H_0 : \beta_2 = \beta_3 = 0$. We evaluate the power in a sequence of alternatives with parameters $(\beta_1, \beta_2, \beta_3) = (2, 2\theta, \theta)$ for each given θ.

For $\theta = 0$, the alternative hypothesis becomes the null hypothesis. According to Theorems 3.3 and 4.2, the distribution of the PLR and Wald statistics should be asymptotically χ^2_2. To verify this empirically, we plot the quantiles of the 1000 PLR statistics against the quantiles of χ^2_2. Figure 1(a) shows the Q-Q plot for case (i). The plot depicts the PLR statistic closely following the χ^2_2 distribution, which is consistent with our asymptotic theory. Figure 1(b) depicts the Q-Q plot for the Wald statistic.

To evaluate the power of the hypothesis test more accurately, we also use the conditional bootstrap method to calculate the critical value for the PLR and Wald tests. In this case, let $\{\hat{\alpha}(\cdot), \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$ be the estimate under the alternative hypothesis. For each simulation, we generate another 1000 bootstrap samples from the model

$$Y_i^* = \hat{\alpha}(U_i) + \hat{\beta}_1 Z_{i1} + \varepsilon_i^*, \qquad i = 1, \ldots, n,$$

where $\varepsilon^* \sim N(0, \hat{\sigma}^2)$. Based on $\{(U_i, Z_{i1}, Z_{i2}, Z_{i3}, Y_i{}^*), i = 1, \ldots, n\}$, compute the PLR statistics and the Wald statistics, and use the 99th, 95th and 90th percentiles as the critical values for testing at the significance levels of 0.01, 0.05 and 0.1, respectively. This method is valid due to the Wilks type phenomenon.

Figure 2(a) depicts the power functions of the PLR test based on 1000 simulations of sample size 100 at three different significance levels: 0.01, 0.05, and 0.1. By using the conditional bootstrap method, the powers at $\theta = 0$ for the above three significance levels are 0.009, 0.042, and 0.101, respectively. This shows that the conditional bootstrap method gives the right level of testing. The power functions increase rapidly as θ increases. This in turn shows that the PLR statistic proposed in Section 3.1 works well. Figures 2(b) and 2(c) depict the simulated power functions of the Wald test and the *F*-test.

In additional to the conditional bootstrap method, the χ^2-approximation can be employed to determine the critical value. The power functions of this method are shown by the dash-

Table 1. Means and standard deviations of estimators for case (i), Sample Size = 100

		$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$	
θ	h	Mean(SD_{m})	$SD(SD_{\mathrm{std}})$	Mean(SD_{m})	$SD(SD_{\mathrm{std}})$	Mean(SD_{m})	$SD(SD_{\mathrm{std}})$
0	DBE	2.0058(0.1376)	0.1348(0.0157)	0.0003(0.1348)	0.1339(0.0145)	−0.0064(0.2263)	0.2224(0.0188)
	0.4167	2.0038(0.1401)	0.1291(0.0153)	−0.0016(0.1348)	0.1285(0.0145)	0.0026(0.2306)	0.2134(0.0188)
	0.625	2.0035(0.1382)	0.1296(0.0152)	−0.0013(0.1335)	0.1290(0.0143)	0.0013(0.2274)	0.2144(0.0184)
	0.9375	2.0039(0.1374)	0.1313(0.0152)	−0.0008(0.1327)	0.1306(0.0142)	0.0004(0.2255)	0.2170(0.0181)
0.5	DBE	2.0014(0.1398)	0.1342(0.0147)	1.0010(0.1392)	0.1346(0.0141)	0.5057(0.2295)	0.2225(0.0192)
	0.4167	2.0066(0.1387)	0.1282(0.0149	0.0000(0.1438)	0.1288(0.0148)	0.5081(0.2362)	0.2125(0.0195)
	0.625	2.0064(0.1375)	0.1288(0.0147)	1.0001(0.1418)	0.1294(0.0146)	0.5096(0.2320)	0.2135(0.0191)
	0.9375	2.0045(0.1368)	0.1304(0.0146)	1.0007(0.1412)	0.1310(0.0145)	0.5100(0.2299)	0.2162(0.0189)

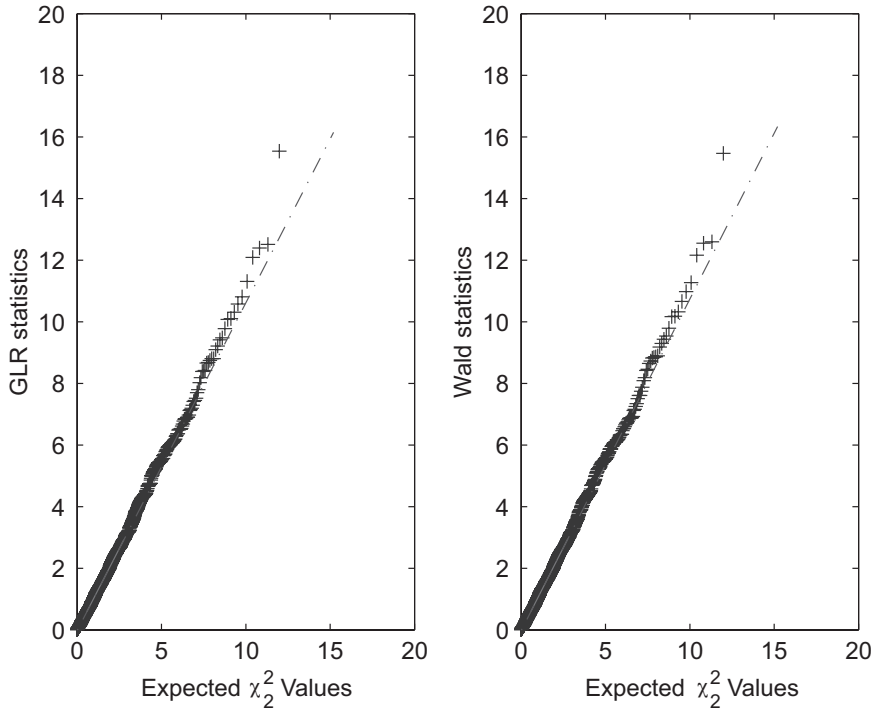

Figure 1. *Q-Q* plots for case (i): (a) PLR statistics; (b) Wald statistics.

dotted curves in Figure 2. The sizes of the test are close to the significance level, keeping in mind that the Monte Carlo error is of size $\sqrt{0.05 \times 0.95/1000} \approx 0.7\%$ at the 5% significance level. However, there is a small upward bias, which is due partially to the bias in the estimation of the parametric component. The non-normal error of case (ii) yields similar results. However, to save space, we omit the presentation.

7.2. A varying-coefficient partially linear model

Simulation data are generated from the varying-coefficient partially linear model

$$Y = \sin(6\pi U)X_1 + \sin(2\pi U)X_2 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \varepsilon,$$

where U, X_1, X_2, Z_1, Z_2, Z_3 are covariates. The covariate U is uniformly distributed on [0, 1]. The covariates X_1, X_2, Z_1, Z_2 are jointly normally distributed with mean 0 and variance 1. Furthermore, the correlation coefficients among these four random variables are $\frac{2}{3}$. The covariate Z_3 is binary and takes the value 1 with probability 0.4. The noise ε is normally distributed with mean 0 and variance 1. In addition, U, (X_1, X_2, Z_1, Z_2), Z_3, and ϵ are

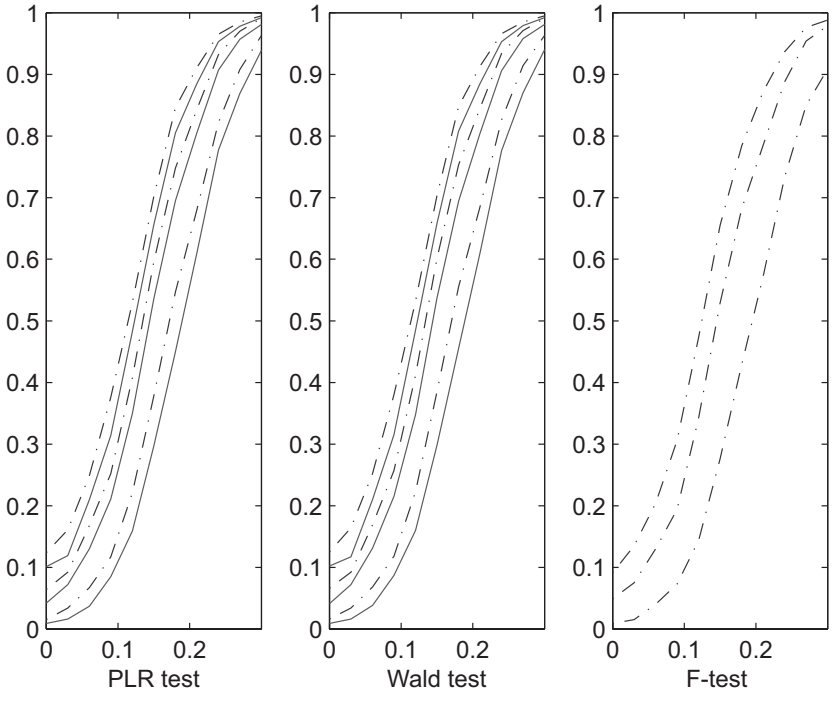

Figure 2. The simulated power functions for case (i) with $h = 0.625$. The critical values are computed by χ^2_2-approximation (dash-dotted lines) and the conditional bootstrap method (solid lines), (a) PLR test, (b) Wald test, (c) F-test.

simulated independently. The true parameter for β_1 is always fixed at $\beta_1 = 2$, and β_2 and β_3 are taken differently for different problems.

For this example, we draw 1000 random samples of size 100 from the above model. Through the cross-validation method, $h = 0.25$ is chosen as the smoothing parameter. Table 2 shows that the performance of the profile least-squares estimator does not sensitively depend on the choice of the bandwidth. Furthermore, the standard error formulae work very well. Figure 3 demonstrates that the PLR and Wald statistics follow the χ^2_2 distribution closely, though there are some biases in the right-hand tail.

We consider the null hypothesis, $H_0 : \beta_2 = \beta_3 = 0$. We evaluate the power of the PLR and Wald tests in a sequence of alternatives with parameters $(\beta_1, \beta_2, \beta_3) = (2, 2\theta, \theta)$ for each given θ. Figure 4 summarizes the result for $h = 0.25$. Again, there is some upward bias for the χ^2-approximation.

7.3. Application to Boston housing data

We now illustrate the proposed method by an application to the Boston housing data set. The data set consists of the median value of owner-occupied homes in 506 US census tracts

Table 2. Means and standard deviations of the estimators for the model of Section 7.2

		$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$	
θ	h	$Mean(SD_m)$	$SD(SD_{std})$	$Mean(SD_m)$	$SD(SD_{std})$	$Mean(SD_m)$	$SD(SD_{std})$
0	0.166	1.9941(0.1779)	0.1643(0.0178)	0.0002(0.1792)	0.1639(0.0165)	0.0065(0.1851)	0.1694(0.0167)
	0.25	1.9947(0.1928)	0.1775(0.0197)	0.0019(0.1941)	0.1772(0.0182)	0.0026(0.2021)	0.1831(0.0183)
	0.375	1.9951(0.1973)	0.1832(0.0203)	0.0023(0.1984)	0.1828(0.0190)	0.0014(0.2053)	0.1889(0.0191)
0.5	0.166	1.995(0.1760)	0.1637(0.0173)	0.9996(0.1806)	0.1629(0.0176)	0.5087(0.1797)	0.1692(0.0173)
	0.25	2.0002(0.1950)	0.1770(0.0189)	1.0000(0.1959)	0.1762(0.0198)	0.5100(0.1964)	0.1830(0.0193)
	0.375	1.9988(0.2019)	0.1825(0.0195)	1.0023(0.2005)	0.1818(0.0205)	0.5112(0.2013)	0.1888(0.0200)

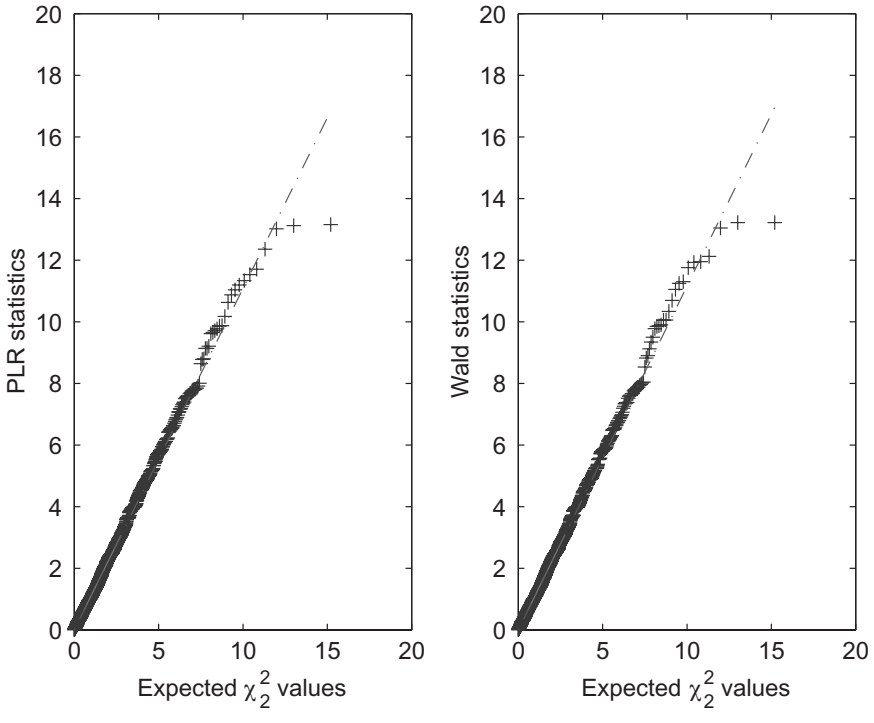

Figure 3. *Q-Q* plots, $h = 0.25$. (a) PLR statistics. (b) Wald statistics.

in the Boston area in 1970, as well as several variables which might explain the variation in housing value (see Harrison and Rubinfeld, 1978). Seven variables, CRIM (per-capita crime rate by town), RM (average number of rooms per dwelling), TAX (full-value property-tax rate per \$10 000), NOX (nitric oxide concentration in parts per 10 million), PTRATIO (pupil–teacher ratio by town), AGE (proportion of owner-occupied units built prior to 1940), and LSTAT (percentage of lower income status of the population) are considered here. For simplicity of notation, the covariates CRIM, RM, TAX, NOX, PTRATIO, and AGE are denoted respectively by $\mathbf{X}_2, \ldots, \mathbf{X}_7$. The objective of the study is to understand the association between the median value of owner-occupied homes and the seven covariates. For comparison, we fitted the multiple linear regression using the seven independent variables. The multiple R^2 is 0.7212, and the residual standard deviation is $\hat{\sigma} = 4.8514$.

We take $\mathbf{X}_1 = 1$ as the intercept term and $U = \sqrt{\text{LSTAT}}$. This allows us to fit a different linear model for a different percentage of a lower income status of the population and permits us to examine how it interacts with other independent variables. Examination of the distribution of LSTAT reveals that it is asymmetric. Thus, the square-root transformation is employed and the resulting data have nearly symmetric distribution. This transform does not

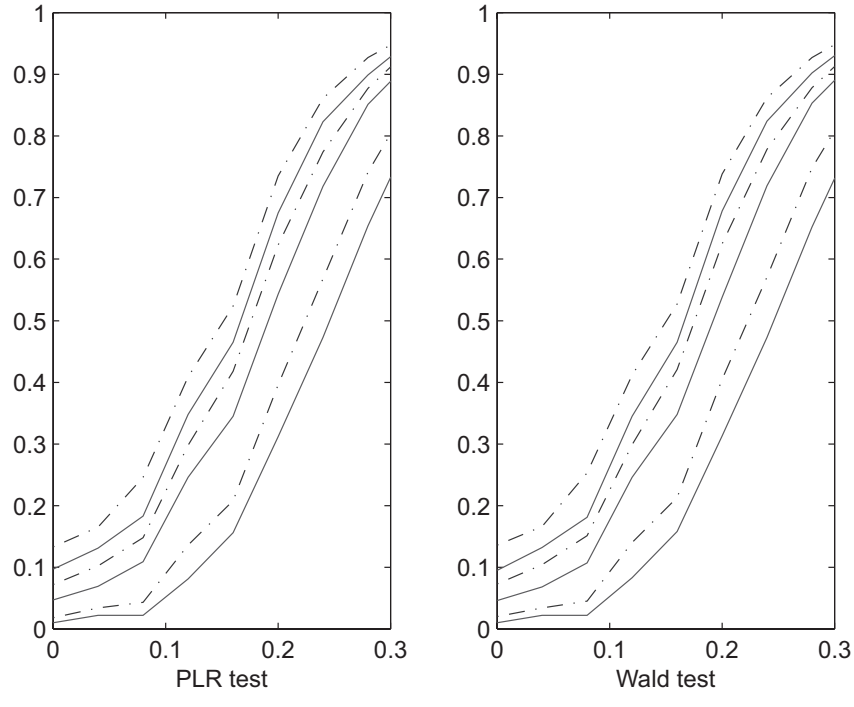

Figure 4. The simulated power functions with $h = 0.625$. *The critical values are computed by χ^2_2-approximation (dash-dotted lines) and the conditional bootstrap method (solid lines). (a) PLR test, (b) Wald test.*

alter the model, but it facilitates our implementation. The Epanechnikov kernel is employed, and the bandwidth is chosen to be 25% of the interval length ($h = 1.2117$).

First, the varying-coefficient model

$$\mathbf{Y} = a_1(U) + \sum_{i=2}^{7} a_i(U)\mathbf{X}_i + \varepsilon$$

is fitted to the given data. The concept of multiple R^2 can be extended to the current context. It is defined as $1 - \mathrm{RSS}/\sum_i (Y_i - \bar{Y})^2$, where RSS is the residual sum of squares. For this model, the multiple R^2 is 0.8345 and the residual standard deviation is 3.7378. To examine the extent to which the association varies over U, we apply the GLR test to see whether each coefficient function is statistically significant. Table 3 presents the p-value for each testing problem, and shows that variables AGE and PTRATIO are not significant at level 1%.

Based on the above analysis, we set the coefficients of AGE and PTRATIO to be constants, and employ the varying-coefficient partially linear model,

$$\mathbf{Y} = a_1(U) + a_2(U)\mathbf{X}_2 + a_3(U)\mathbf{X}_3 + a_4(U)\mathbf{X}_4 + a_5(U)\mathbf{X}_5 + b_1\mathbf{X}_6 + b_2\mathbf{X}_7 + \varepsilon,$$

Table 3. p-values for testing whether a coefficient functions is constant

	$a_1(U)$	$a_2(U)$	$a_3(U)$	$a_4(U)$	$a_5(U)$	$a_6(U)$	$a_7(U)$
GLR statistics	10.3173	30.4889	27.1646	11.3421	7.3747	4.4777	1.5513
p-values	0.0002	0.0000	0.0000	0.0001	0.0030	0.0441	0.4815

Table 4. p-values for testing whether a coefficient function is zero

	b_1	b_2		b_1	b_2
PLR statistics	24.3115	0.0752	Wald statistics	51.3412	0.1515
p-values	0.0000	0.6971	p-values	0.0000	0.6981

to fit the given data. A natural question is whether the coefficients of AGE and PTRATIO are statistically significant. To answer this question, the proposed PLR and Wald tests are employed. The p-values for the tests are summarized in Table 4, which indicates that the coefficient of AGE is zero.

Finally, the varying-coefficient partially linear model

$$\mathbf{Y} = a_1(U) + a_2(U)\mathbf{X}_2 + a_3(U)\mathbf{X}_3 + a_4(U)\mathbf{X}_4 + a_5(U)\mathbf{X}_5 + b_1\mathbf{X}_6 + \varepsilon$$

is fitted to the given data. The estimated parametric coefficient is $b_1 = -0.7199$ with an estimated standard error of 0.0998. Figure 5 depicts the coefficient functions. The multiple R^2 is 0.8304 and the residual standard deviation is 3.7836. The result shows that in the tracts with crowded schools, the value of housing tends to be lower.

Appendix

We outline the key idea of the proof. The following technical conditions are imposed. They are not the weakest possible conditions, but they are imposed to facilitate the technical proofs.

(1) The random variable U has a bounded support $\mathbf{\Omega}$. Its density function $f(\cdot)$ is Lipschitz continuous and bounded away from 0 on its support.
(2) The $k \times k$ matrix $\mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U)$ is non-singular for each $U \in \mathbf{\Omega}$. $\mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U)$, $\mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U)^{-1}$ and $\mathrm{E}(\mathbf{X}\mathbf{Z}^{\mathrm{T}}|U)$ are all Lipschitz continuous.
(3) There is an $s > 2$ such that $\mathrm{E}\|\mathbf{X}\|^{2s} < \infty$ and $\mathrm{E}\|\mathbf{Z}\|^{2s} < \infty$ and for some $\varepsilon < 2 - s^{-1}$ such that $n^{2\varepsilon-1}h \to \infty$.
(4) $\{\alpha_i(\cdot), i = 1, \ldots, p\}$ have continuous second derivative in $U \in \mathbf{\Omega}$.
(5) The function $K(\cdot)$ is a symmetric density function with compact support.
(6) $nh^8 \to 0$ and $nh^2/(\log n)^2 \to \infty$.

The following notation will be used in the proof of the lemmas and theorems.

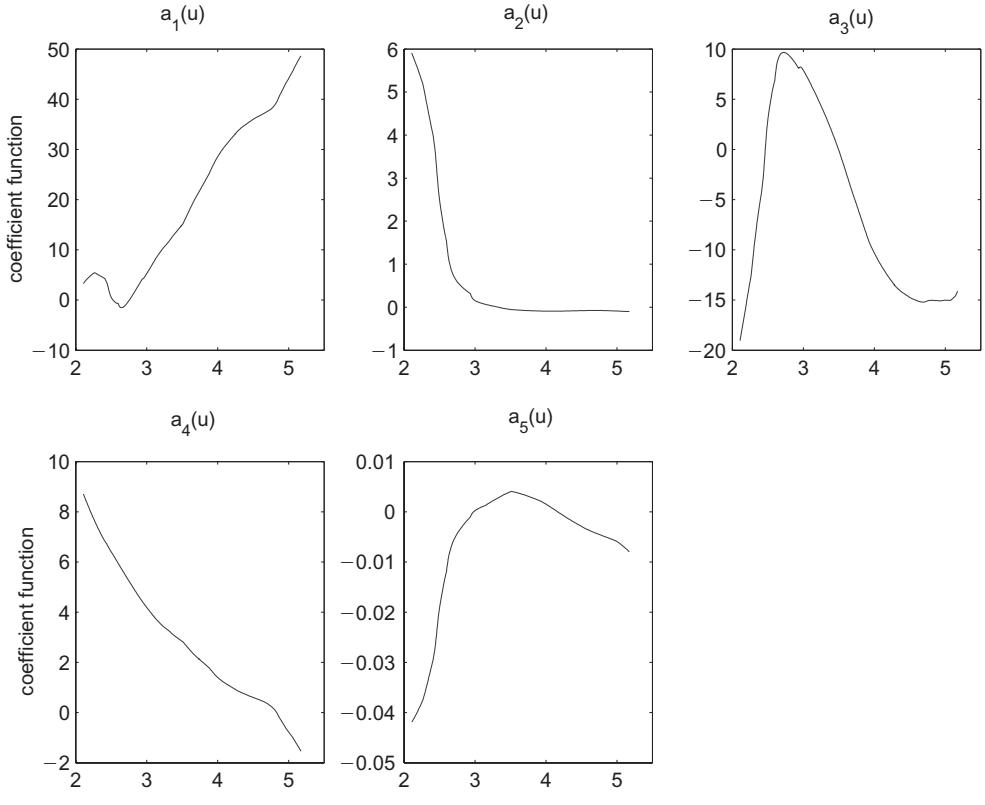

Figure 5. The estimated coefficient functions.

Let $\mu_i = \int u^i K(u)\mathrm{d}u$, $\nu_i = \int u^i K^2(u)\mathrm{d}u$ and $c_n = \{\log(1/h)/nh\}^{1/2} + h^2$. Set $\Gamma(U) = \mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U)$, $\Phi(U) = \mathrm{E}(\mathbf{X}\mathbf{Z}^{\mathrm{T}}|U)$.

Lemma A.1. *Let* $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ *be independent and identically distributed random vectors, where the* $\mathbf{Y}_i$ *are scalar random variables. Further assume that* $\mathrm{E}|y|^s < \infty$ *and* $\sup_x \int |y|^s f(x, y)\mathrm{d}y < \infty$, *where* f *denotes the joint density of* $(\mathbf{X}, \mathbf{Y})$. *Let* K *be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Given that* $n^{2\varepsilon-1}h \to \infty$ *for some* $\varepsilon < 1 - s^{-1}$, *then*

$$\sup_x \left| \frac{1}{n}\sum_{i=1}^{n} [K_h(\mathbf{X}_i - x)\mathbf{Y}_i - \mathrm{E}\{K_h(\mathbf{X}_i - x)\mathbf{Y}_i\}] \right| = O_p\left(\left\{ \frac{\log(1/h)}{nh} \right\}^{1/2} \right).$$

Proof. This follows immediately from the result obtained by Mack and Silverman (1982).

Lemma A.2. *Under conditions* (1)–(6), *we have*

$$n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}} \xrightarrow{\mathrm{P}} \mathrm{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) - \mathrm{E}\big[\mathrm{E}(\mathbf{Z}\mathbf{X}^{\mathrm{T}}|U)\mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U)^{-1}\mathrm{E}(\mathbf{X}\mathbf{Z}^{\mathrm{T}}|U)\big].$$

Furthermore, $\hat{\mathbf{\Sigma}}_h^* \xrightarrow{\mathrm{P}} \mathbf{\Sigma}$.

Proof. In equation (2.4), observe that the smoothing matrix $\mathbf{S}$ has the form

$$\mathbf{S} = \begin{pmatrix} [\mathbf{X}_1^{\mathrm{T}} \quad 0]\{\mathbf{D}_{u_1}^{\mathrm{T}}\mathbf{W}_{u_1}\mathbf{D}_{u_1}\}^{-1}\mathbf{D}_{u_1}^{\mathrm{T}}\mathbf{W}_{u_1} \\ \vdots \\ [\mathbf{X}_n^{\mathrm{T}} \quad 0]\{\mathbf{D}_{u_n}^{\mathrm{T}}\mathbf{W}_{u_n}\mathbf{D}_{u_n}\}^{-1}\mathbf{D}_{u_n}^{\mathrm{T}}\mathbf{W}_{u_n} \end{pmatrix}.$$

Note that

$$\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{D}_u = \begin{pmatrix} \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}K_h(U_i-U) & \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\left(\frac{U_i-U}{h}\right)K_h(U_i-U) \\ \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\left(\frac{U_i-U}{h}\right)K_h(U_i-U) & \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\left(\frac{U_i-U}{h}\right)^2K_h(U_i-U) \end{pmatrix}.$$

Each element of the above matrix is in the form of a kernel regression. By Lemma A.1,

$$\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{D}_u = nf(U)\Gamma(U) \otimes \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix} \{1 + O_P(c_n)\} \tag{A.1}$$

holds uniformly in U, where $\otimes$ is the Kronecker product. By the same argument,

$$\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{Z} = nf(U)\Phi(U) \otimes (1, 0)^{\mathrm{T}}\{1 + O_P(c_n)\} \tag{A.2}$$

holds uniformly in U. Combining the last two results yields that, uniformly in $U \in \mathbf{\Omega}$,

$$[\mathbf{X}^{\mathrm{T}}, 0]\{\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{D}_u\}^{-1}\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{Z} = \mathbf{X}^{\mathrm{T}}\Gamma(U)^{-1}\Phi(U)\{1 + O_p(c_n)\}. \tag{A.3}$$

Equivalently, we have

$$\mathbf{S}\mathbf{Z} = \begin{pmatrix} \mathbf{X}_1^{\mathrm{T}}\Gamma(U_1)^{-1}\Phi(U_1) \\ \vdots \\ \mathbf{X}_n^{\mathrm{T}}\Gamma(U_n)^{-1}\Phi(U_n) \end{pmatrix}\big\{1 + O_p(c_n)\big\}.$$

Now, using (A.3) and some algebra, it is easy to show that

$$n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}} = n^{-1}\sum_{i=1}^n \big[\mathbf{Z}_i - \Phi(U_i)^{\mathrm{T}}\Gamma(U_i)^{-1}\mathbf{X}_i\big]\big[\mathbf{Z}_i^{\mathrm{T}} - \mathbf{X}_i^{\mathrm{T}}\Gamma(U_i)^{-1}\Phi(U_i)\big]\{1 + O_p(c_n)\}.$$

By the law of large numbers, the result holds. □

Lemma A.3. *Under conditions* (1)–(6)*, we have*

$$n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I}-\mathbf{S})(\mathbf{I}-\mathbf{S})^{\mathrm{T}}\widetilde{\mathbf{Z}} \xrightarrow{\mathrm{P}} \mathrm{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}}) - \mathrm{E}\big[\mathrm{E}(\mathbf{Z}\mathbf{X}^{\mathrm{T}}|U)\mathrm{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}|U)^{-1}\mathrm{E}(\mathbf{X}\mathbf{Z}^{\mathrm{T}}|U)\big].$$

Furthermore, $\hat{\mathbf{\Sigma}}_h \xrightarrow{\mathrm{P}} \mathbf{\Sigma}$.

Proof. Observe that

$$n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I}-\mathbf{S})(\mathbf{I}-\mathbf{S})^{\mathrm{T}}\widetilde{\mathbf{Z}} = n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}} - J_1 - J_2 + J_3,$$

where $J_1 = n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\mathbf{S}\widetilde{\mathbf{Z}}$, $J_2 = n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\widetilde{\mathbf{Z}}$, $J_3 = n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\mathbf{S}\mathbf{S}^{\mathrm{T}}\widetilde{\mathbf{Z}}$. The result follows from Lemma A.2, if we can show that J_1, J_2 and J_3 are of order $o_P(1)$. Set $\widetilde{\mathbf{Z}} = (\widetilde{\mathbf{Z}}_1, \ldots, \widetilde{\mathbf{Z}}_n)^{\mathrm{T}}$. Then

$$J_1 = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i[\mathbf{X}_i^{\mathrm{T}} \quad 0]\{\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\mathbf{D}_{u_i}\}^{-1}\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\widetilde{\mathbf{Z}}.$$

By using the same argument that leads to (A.2), we obtain

$$\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\widetilde{\mathbf{Z}} = nf(U)\Phi(U)\otimes(1 \qquad 0)^{\mathrm{T}}O_p(c_n).$$

Combining this with (A.1) yields

$$[\mathbf{X}^{\mathrm{T}}, 0]\{\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{D}_u\}^{-1}\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\widetilde{\mathbf{Z}} = \mathbf{X}^{\mathrm{T}}\Gamma(U)^{-1}\Phi(U)O_p(c_n). \tag{A.4}$$

Hence,

$$J_1 = n^{-1}\sum_{i=1}^{n}\big[\mathbf{Z}_i - \Phi(U_i)^{\mathrm{T}}\Gamma(U_i)^{-1}\mathbf{X}_i\{1+O_p(c_n)\}\big]\mathbf{X}_i^{\mathrm{T}}\Gamma(U_i)^{-1}\Phi(U_i)O_p(c_n).$$

Note that by the law of large numbers, the main term cancels. Hence, by applying the central limit theorem, we have $J_1 = O_p(c_n^2)$. Analogously, we can show that $J_2 = O_p(c_n^2)$ and $J_3 = O_p(c_n^2)$. □

Lemma A.4. *Under conditions* (1)–(6)*, we have*

$$n^{-1}\widetilde{\mathbf{Z}}(\mathbf{I}-\mathbf{S})\mathbf{M} = O_P(c_n^2).$$

Proof. Observe that

$$n^{-1}\widetilde{\mathbf{Z}}(\mathbf{I}-\mathbf{S})\mathbf{M} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i\{\mathbf{X}_i^{\mathrm{T}}\alpha(U_i) - [\mathbf{X}_i^{\mathrm{T}} \quad 0]\{\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\mathbf{D}_{u_i}\}^{-1}\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\mathbf{M}\}.$$

Similarly to (A.4), we can show that the following equation holds uniformly in $U \in \boldsymbol{\Omega}$:

$$[\mathbf{X}^{\mathrm{T}} \quad 0]\{\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{D}_u\}^{-1}\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{M} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\alpha}(U)\{1+O_p(c_n)\}.$$

Also, by (A.3), we have $\widetilde{\mathbf{Z}}_i^{\mathrm{T}} = \mathbf{Z}_i^{\mathrm{T}} - \mathbf{X}_i^{\mathrm{T}}\Gamma(U_i)^{-1}\Phi(U_i)\{1+O_p(c_n)\}$, $i = 1, \ldots, n$. Hence,

$$\begin{aligned} n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I}-\mathbf{S})\mathbf{M} &= n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i\{\mathbf{X}_i^{\mathrm{T}}\alpha(U_i) - [\mathbf{X}_i^{\mathrm{T}} \quad 0]\{\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\mathbf{D}_{u_i}\}^{-1}\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\mathbf{M}\} \\ &= n^{-1}\sum_{i=1}^{n}\big[\mathbf{Z}_i - \Phi(U_i)^{\mathrm{T}}\Gamma(U_i)^{-1}\mathbf{X}_i\big]\mathbf{X}_i^{\mathrm{T}}\alpha(U_i)\{1+O_p(c_n)\}O_p(c_n) \\ &= O_p(c_n^2). \end{aligned}$$

□

Proof of Theorem 4.1. By (2.6), we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})(\mathbf{M} + \boldsymbol{\varepsilon}).$$

By Lemmas A.2 and A.4, the bias term

$$\sqrt{n}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\mathbf{M} = O_p(\sqrt{n}c_n^2).$$

Consider the stochastic term. By Lemma A.2, we have

$$n^{1/2}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon} = n^{-1/2}\sigma^{-2}\boldsymbol{\Sigma}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon}\{1 + o_P(1)\}. \tag{A.5}$$

Note that,

$$\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon} = \sum_{i=1}^{n}\widetilde{\mathbf{Z}}_i\{\varepsilon_i - [\mathbf{X}_i^{\mathrm{T}}, 0]\{\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\mathbf{D}_{u_i}\}^{-1}\mathbf{D}_{u_i}^{\mathrm{T}}\mathbf{W}_{u_i}\boldsymbol{\varepsilon}\}.$$

By using the same argument as before, we have

$$[\mathbf{X}^{\mathrm{T}}, 0]\{\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\mathbf{D}_u\}^{-1}\mathbf{D}_u^{\mathrm{T}}\mathbf{W}_u\boldsymbol{\varepsilon} = \mathbf{X}^{\mathrm{T}}\Gamma(U)^{-1}\mathrm{E}(\mathbf{X}|U)O_p(c_n).$$

Then we can show that

$$\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon} = \sum_{i=1}^{n}\{\mathbf{Z}_i - \Phi(U_i)^{\mathrm{T}}\Gamma(U_i)^{-1}\mathbf{X}_i\}\varepsilon_i\{1 + o_P(1)\}.$$

By the Slutsky theorem and the central limit theorem, we have

$$n^{-1/2}\sigma^{-2}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon} \to N(0, \boldsymbol{\Sigma}^{-1}).$$

This, together with (A.5), proves the result. □

Proof of Theorems 3.1 and 3.2. Theorem 3.1 is a specific case of Theorem 3.2. First, we show that $n^{-1}\mathrm{RSS}_1 = \sigma^2(1 + o_p(1))$. By (2.6) and (2.7), we have

$$\begin{aligned}
\mathrm{RSS}_1 &= \sum_{i=1}^{n}(Y_i - \hat{\mathbf{M}}_i - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{Z}_i)^2 = [\mathbf{Y} - \hat{\mathbf{M}} - \mathbf{Z}\hat{\boldsymbol{\beta}}]^{\mathrm{T}}[\mathbf{Y} - \hat{\mathbf{M}} - \mathbf{Z}\hat{\boldsymbol{\beta}}] \\
&= [\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}]^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})[\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}] \\
&= [\mathbf{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{M} + \boldsymbol{\varepsilon}]^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})[\mathbf{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \mathbf{M} + \boldsymbol{\varepsilon}] \\
&= I_1 + I_2 + I_3 + I_4 + I_5 + I_6,
\end{aligned}$$

where

$$I_1 = \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon}, \quad I_4 = \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\mathbf{M} + \mathbf{M}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon},$$

$$I_2 = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \quad I_5 = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\mathbf{M} + \mathbf{M}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}\widetilde{\mathbf{Z}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

$$I_3 = \mathbf{M}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\mathbf{M}, \quad I_6 = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\widetilde{\mathbf{Z}}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathrm{T}}(\mathbf{I} - \mathbf{S})^{\mathrm{T}}\widetilde{\mathbf{Z}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

By using the same argument as before, it can be shown that

$$n^{-1}I_1 = \sigma^2\{1 + o_p(1)\}, \qquad n^{-1}I_2 = O_p(n^{-1}), \qquad n^{-1}I_3 = O_p(c_n^2),$$
$$n^{-1}I_4 = O_p(c_n), \qquad n^{-1}I_5 = O_p(n^{-1/2}c_n), \qquad n^{-1}I_6 = O_p(n^{-1/2}).$$

Similarly, RSS_0 can be decomposed as

$$\begin{aligned}\mathrm{RSS}_0 &= [\mathbf{Y} - \hat{\mathbf{M}}_0 - \mathbf{Z}\hat{\boldsymbol{\beta}}_0]^{\mathrm{T}}[\mathbf{Y} - \hat{\mathbf{M}}_0 - \mathbf{Z}\hat{\boldsymbol{\beta}}_0] \\ &= [\mathbf{Y} - \hat{\mathbf{M}} - \mathbf{Z}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{S})\mathbf{Z}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)]^{\mathrm{T}}[\mathbf{Y} - \hat{\mathbf{M}} - \mathbf{Z}\hat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{S})\mathbf{Z}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)] \\ &= \mathrm{RSS}_1 + J_1 + J_2 + J_3,\end{aligned}$$

where

$$J_1 = [(\mathbf{I} - \mathbf{S})\mathbf{Z}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)]^{\mathrm{T}}[(\mathbf{I} - \mathbf{S})\mathbf{Z}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)],$$

$$J_2 = [\mathbf{Y} - \hat{\mathbf{M}} - \mathbf{Z}\hat{\boldsymbol{\beta}}]^{\mathrm{T}}[(\mathbf{I} - \mathbf{S})\mathbf{Z}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)],$$

$$J_3 = [(\mathbf{I} - \mathbf{S})\mathbf{Z}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)]^{\mathrm{T}}[\mathbf{Y} - \hat{\mathbf{M}} - \mathbf{Z}\hat{\boldsymbol{\beta}}].$$

As the estimators for $\boldsymbol{\beta}$ under the null and alternative hypotheses have the relation

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} - (\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\mathbf{A}^{\mathrm{T}}\{\mathbf{A}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\mathbf{A}^{\mathrm{T}}\}^{-1}\mathbf{A}\hat{\boldsymbol{\beta}},$$

J_1 can be written as

$$J_1 = [\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)]^{\mathrm{T}}[\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)] = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\{\mathbf{A}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\mathbf{A}^{\mathrm{T}}\}^{-1}\mathbf{A}\hat{\boldsymbol{\beta}}.$$

By Lemma A.3, we have $n^{-1}\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}} \to \boldsymbol{\Sigma}/\sigma^2$. This, together with the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and the proofs of Theorems 4.2 and 4.3, gives

$$J_1 = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\{\mathbf{A}(\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}})^{-1}\mathbf{A}^{\mathrm{T}}\}^{-1}\mathbf{A}\hat{\boldsymbol{\beta}}^{\mathrm{T}} \xrightarrow{\mathrm{P}} \sigma^2\chi_l^2(\lambda).$$

It is easy to show that $J_2 = J_3 = 0$. Thus,

$$\mathrm{RSS}_0 - \mathrm{RSS}_1 \xrightarrow{\mathrm{P}} \sigma^2\chi_l^2(\lambda).$$

Then, by the Slutsky theorem,

$$2T_n(h) = n\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1} \xrightarrow{\mathrm{P}} \chi_l^2(\lambda). \qquad \square$$

Proof of Theorems 4.2 and 4.3. Theorem 4.2 is a specific case of Theorem 4.3. By Theorem 4.1, it is easy to see that

$$\sqrt{n}\mathbf{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}).$$

By Lemma A.3 that $\hat{\boldsymbol{\Sigma}}_h \xrightarrow{\mathrm{P}} \boldsymbol{\Sigma}$ and the Slutsky theorem, then

$$W_n(h) = \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\hat{\boldsymbol{\Sigma}}_h\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{P}} \chi^2_l(\lambda). \qquad \square$$

Proof of Theorem 5.1. Let $\boldsymbol{\beta}$ be the true parameter, and $\mathbf{Y}^* = \mathbf{Y} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}$. We then transform the semiparametric model (1.1) to the nonparametric varying-coefficient model

$$\mathbf{Y}^* = \alpha_1(\cdot)\mathbf{X}_1 + \cdots + \alpha_p(\cdot)\mathbf{X}_p + \varepsilon.$$

Analogously, we can define the GLR statistic for problem (5.1) as

$$T_0^* = \frac{n}{2}\log\frac{\mathrm{RSS}^*(H_0)}{\mathrm{RSS}^*(H_1)},$$

where $\mathrm{RSS}^*(H_0) = \sum_{i=1}^n (Y_i^* - \sum_{j=1}^p \widetilde{\alpha}_j^* X_{ij})^2$ and $\mathrm{RSS}^*(H_1) = \sum_{i=1}^n (Y_i^* - \sum_{j=1}^p \hat{\alpha}_j^*(U_i)X_{ij})^2$. By Theorem 5 of Fan *et al.* (2001), we have $r_K T_0^* \overset{a}{\sim} \chi^2_{\delta_n}$. The proof will be complete by showing the following two claims: (i) $n^{-1}\{\mathrm{RSS}(H_0) - \mathrm{RSS}^*(H_0)\} = o_P(1)$; (ii) $n^{-1}\{\mathrm{RSS}(H_1) - \mathrm{RSS}^*(H_1)\} = o_P(1)$. Claim (i) follows from the fact that

$$\mathrm{RSS}^*(H_0) = \sigma^2(n-p)\{1 + o_P(1)\} \quad \text{and} \quad \mathrm{RSS}(H_0) = \sigma^2(n-p-q)\{1 + o_P(1)\}.$$

Claim (ii) follows from the fact that

$$\mathrm{RSS}^*(H_1) - \mathrm{RSS}(H_1) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sigma^2 q\{1 + o_P(1)\} = o_P(n). \qquad \square$$

Acknowledgements

This research was partially supported by National Science Foundation grants DMS-0355179 and DMS-0354223, National Institutes of Health grant R01 HL69720, and Research Grants Council grant CUHK 4262/01P of the Hong Kong Special Administrative Region.

References

Bickel, P.J. and Kwon, J. (2001) Inference for semiparametric models: Some current frontiers (with discussion). *Statist. Sinica*, **11**, 863–960.

Bickel, P.J., Klaassen, A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and Adaptive Inference in Semi-parametric Models*. Baltimore, MD: Johns Hopkins University Press.

Brumback, B. and Rice, J.A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.*, **93**, 961–994.

Cai, Z., Fan, J. and Li, R. (2000) Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.*, **95**, 888–902.

Carroll, R.J., Fan, J., Gijbels, I, and Wand, M.P. (1997) Generalized partially linear single-index models. *J. Amer. Satist. Assoc.*, **92**, 477–489.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998) Nonparametric estimation via local estimating equations. *J. Amer. Statist. Assoc.*, **93**, 214–227.

Chamberlain, G. (1992) Efficient bounds for semiparametric regression. *Econometrika*, **60**, 567–596.

Chen, R. and Tsay, R.J. (1993) Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.*, **88**, 298–308.

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991) Local regression models. In J.M. Chambers and T.J. Hastie (eds.), *Statistical Models in S*, pp. 309–376. Pacific Grove, CA: Wadsworth/Brooks-Cole.

Cuzick, J. (1992) Semiparametric additive regression. *J. Roy. Statist. Soc. Ser. B*, **54**, 831–843.

Fan, J. and Gijbels, I. (1995) Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaption. *J. Roy. Statist. Soc. Ser. B*, **57**, 371–394.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

Fan, J. and Huang, L. (2001) Goodness-of-fit test for parametric regression models. *J. Amer. Statist. Assoc.*, **96**, 640–652.

Fan, J. and Zhang, W. (1999) Statistical estimation in varying coefficient models. *Ann. Statist.*, **27**, 1491–1518.

Fan, J. and Zhang, W. (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Statist.*, **27**, 715–731.

Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. of Statist.*, **29**, 153–193.

Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

Haggan, V. and Ozaki, T. (1981) Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika*, **68**, 189–196.

Härdle, W., Mammen, E. and Müller, M. (1998) Testing parametric versus semiparametric modelling in generalized linear models. *J. Amer. Statist. Assoc.*, **93**, 1461–1474.

Härdle, W., Liang, H. and Gao, J.T. (2000) *Partially Linear Models*. New York: Springer-Verlag.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (2004) Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, **20**, 265–300.

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *J. Environ. Economics Management*, **5**, 81–102.

Hastie, T.J. and Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman & Hall.

Hastie, T.J. and Tibshirani, R. (1993) Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B*, **55**, 757–796.

Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.-P. (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

Huang, J.Z., Wu, C.O. and Zhou, L. (2002) Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111–128.

Ingster, Yu.I. (1993) Asymptotically minimax hypothesis testing for nonparametric alternatives I–III. *Math. Methods Statist.*, **2**, 85–114; **3**, 171–189; **4**, 249–268.

Li, Q., Huang, C.J., Li, D. and Fu, T.T. (2002) Semiparametric smooth coefficient models. *J. Business Econom. Statist.*, **20**, 412–422.

Liang, H., Härdle, W. and Carroll, R.J. (1999) Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.*, **27**, 1519–1535.

Mack, Y.P. and Silverman, B.W. (1982) Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **61**, 405–415.

Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.*, **92**, 1049–1062.

Ruppert, D., Sheathers, S.J. and Wand, M.P. (1995) An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, **90**, 1257–1270.

Severini, T.A. and Wong, W.H. (1992) Generalized profile likelihood and conditional parametric models. *Ann. Statist.*, **20**, 1768–1802.

Speckman, P. (1988) Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. B*, **50**, 413–436.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman & Hall.

Wahba, G. (1984) Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan–U.S. Joint Seminar, Tokyo, pp. 319–329. Tokyo Institute of Statistical Mathematics.

Xia, Y. and Li, W.K. (1999) On the estimation and testing of functional-coefficient linear models. *Statist. Sinica*, **9**, 735–757.

Yatchew, A. (1997) An elementary estimator for the partial linear model. *Economics Lett.*, **57**, 135–143.

Zhang, W., Lee, S.-Y. and Song, X. (2002) Local polynomial fitting in semivarying coefficient models. *J. Multivariate Anal.*, **82**, 166–188.

Received August 2003 and revised December 2004